DATE: 14-2-2017 SUBJECT: داسشريبه

Big data ⇒ refers of exponential growth and availability of data
"volume" large size - more complex tools - variety - volicity.
valued or not "مفيدة ولا لا" - trusted or not "accurcy of data".
data structure "structured - semi structured - unstructured"

structured	semi structured	unstructured
relational dB	XML Jason Binary Jaison.	video audio PdF organize أجزاء منه ليه

required tools:

- way of sturage.
- analysis tools
- data scientist → personal scales "مشتاك"
  → technical knowledge
  → mathematics tools.

oLaP ⇒ online analatical processing
"التطور لدرجة التخزين" هاك

Slide 17
Island. seperable. "spreadsheets"

1- not secure.
2. replication.
3- hacking.
4-

1- easy and simple.
2-

data ware house.
الغرض الأساسي
هو تخزين البيانات بشكل
آخر أكثر أمانًا

1- controlled.
2- must have permission for apply any action on data.
3- controlled by data Base administrator.

SandBox
بتقدر تاخذ جزء من الـ dWH permission
وبعد ما تأخذ الجزء الى محتاجه من dWH فقط وتنفذ
عليه العمليات المطلوبة

DATE : ______ SUBJECT : ______

وبذلك يبخصوص Highperformance

case study. Bank from local int global.

١- القوانين الخاصة بكل دولة.
٢. طرق التعامل مع البنك.
٣- High security.
٤- تغير العملات.
٥- بزيادة العملاء نحتاج الى عالم زيادة.
٦- ممكن نحتاج تعامل distributed dB.

Data scientist: 1- variety.
2- stream large.
3- managment data.
4-

• Data analisi.

• يشوف data محتاجه أيه وليا عملها نعملها لوحدها

KPI → Keep performance indicator.

Classifier → T.P , FN , TN , Fp

performance evaluation.

*Bussines intelligance: بيستخدم مقاييس المعيش وهو يفسرها بـ as في هذه

ويعمل analysis كل ده علشان أقدر أخد قرار بناء على هذا التحليل.

مش يقدر يتعامل إلا مع "structured data" فقط.

DATE : ____________ SUBJECT : ____________

· Data scientist.

ألترجبرة ولمادة من السابقة
دعقد ، بيتعامل مع أي نوع من data « unstructured _ structured »

Comparison between them.	① scales _ ② Job of each of them	هام ..

33

Throughput ⇒ الك أي حيشتغل
ـه data من شيء لآخر .

---

skales of data scientist — lec2

1- statistices skills
2- Database
3- Critical thinking, creative, Innorate + Communications skill + Adaptive
4- machine learning + Data mining + Advanced mathematics.
5- collect data from differrent online source
6- EXTract data & Analysis
7- can make Correlations & Conneetions.
8- web development & web design.
9- programming skills,

S5

Data enables. ≃ data collectors.
Professional   Business reports, machine learning lim Teal

DATE:
57

• Quantitive — أي يعقد ويقيس الداتا اللي جاية ويعمل Technical report ومعرفة بعض statistics.

. Skeptical — بتكافئ انه يكون بشكاك

. Communications & Collaborative — لديه القدرة على التعامل مع الناس مختلفة skills

5.10

أوبئة pandemics — الأمثلة دي مش مهمة

فلاونزا الخنازير Swine Flu

5.12.

out Come — يبدأ سوف على المرض بـ 5 مراحل بنوعه.

5.13 Life science. " Genome" الصفات الوراثية

A1. ā , Complexity.

A2 .

life cycle of DAP — lec3

1_define problem — 1_ outcomes, 4_ resources available

2_ I/P

3_ issuse.

2_ collect available data and be sure that it's enough and secure.

3_ modeling.

4_ Test model

• يفضل بعد الانتهاء من كل مرحلة التأكد أنه تم تحقيقها بأفضل شكل ممكن.

documentation.

DATE: SUBJECT:

مع كل مرحلة يتم التنبؤ بكل ما تستهدفه الآخر بعد ذلك.
وأيضاً ليه ادخل في عمل additional analysis.
والتأكد من validation للداتا المختلفة

Business user → end user.

S. 9

1) Discovery.
- learn aboute the problem domain.
- hold history of this domain and analysis it
- Documentation of the old project.
- measure to what resource available and period of time and Quality of data.

2- Data pre.
is responsable of building "SandBox".

3. model
1- SW that will uses.
2- Feature important "Feathure selection"
3- HW.

S. 16
probe. إستجوبه

S 20 discussion in next lecture.

DATE: _____ SUBJECT: _____

Big data ⇒ is a popular term which refers to the exponential growth and availability of data, both structured and unstructured.

Three Vs' to describe the definition of big data.

Volume
velocity
variety

Volume.

⇒ There has been a large increase of data volume there reasons:

1. All of the transactional data that has added up over the years.
2. streaming data From social media.
3. machine to machine data increase.

Velocity.

⇒ Data is being streamed at huge speeds and needs To be dealt with a timely manner

1. Social media.
2. mobile devices.

the biggest challenge is how to react Fast enough. To the massive amount of data that is being Flew rapidly.

DATE: _____ SUBJECT: _____

Variety

managing all the different formats is an issue many organizations have to battle.

- there are many different types of data
- Structured.
- Email.
- Audio & video.
- Application data.
- Financial transactions.
- Unstructured documents.
- So to manage many organizations have to battle

Volume, velocity, variety, value, veracity.

Big Data

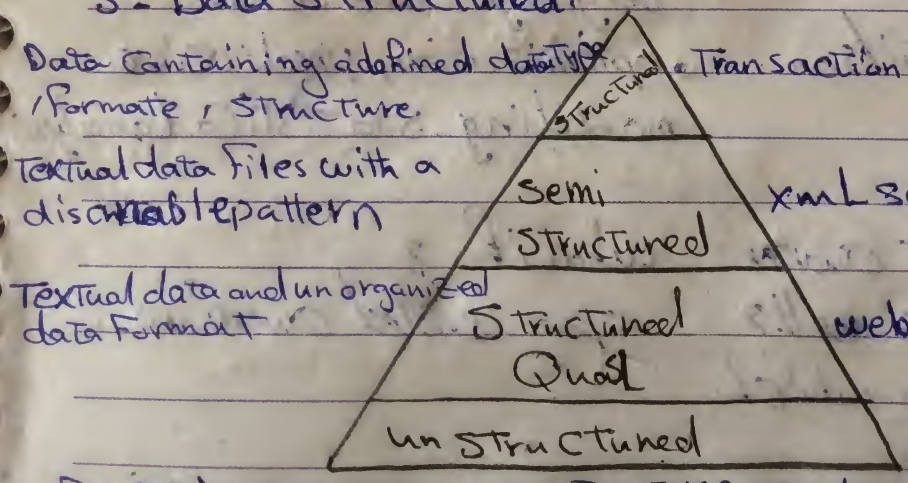
is data whose scale, distribution, diversity, and or timeliness require the use of technical architectures and analytics to enable.

Key characteristics of BD

1- data volume.

2- processing complexity (parallel computing environment and massively parallel processing)

3- Data Structured.



Data containing a defined data type, format, structure. — Transaction data and OLAP

Textual data files with a discernable pattern — xml schema

Textual data and un organized data format — web click stream

ex Text documents, pdf

Data has no inherent structure and in usually stored as different types of files.

DATE: _____ SUBJECT: _____

Storing data

Data islands spreadmarts	Data warehouse	Analytic SandBox
isolated data	Centeralized Data Containers in a purpose-built space	Data assets gethered From multiple Sources and technology For analysis.
• spreedsheet and Low volume DB.	•	• enable high performance
• Analyst dependent on data extracts.	• Analyst dependent on IT @ DBAs For data access and schema change. • Analysts must spened significant time to get extract For multiple sources.	• reduce cas associated with data replicated. • analyist "owned" • more robust analyses.

Business inteligence

Structured data, traditional sources. manageable datasets...

Standard and details on demand.

- what happened last quarter How many did we sell? where is the problem? in which situation.

Data science

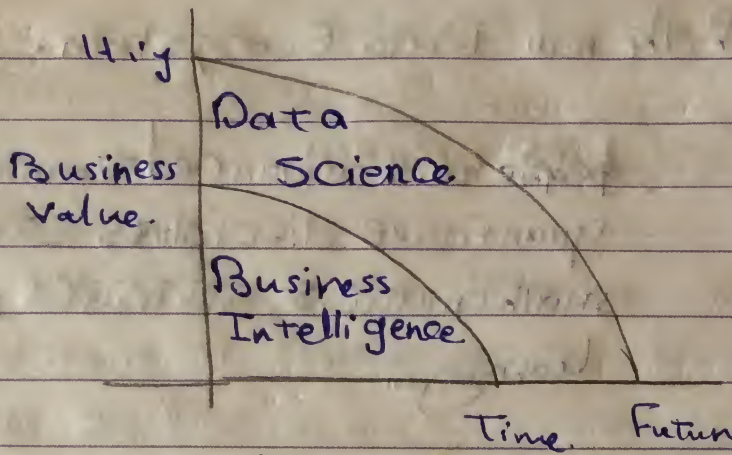
• Structure/unstructured data, many types of sources, very large data sets.

• optimization, predictive modeling. Statistical analysis.

what if..?

open ended questions.

DATE: _____ SUBJECT: _____



implications of typical Architecure For data science.

1- high-value data is hard to reach and leverage.
2. Data is moving in batches From EDW to local analytics.
3. isolated and analytic projects, rather than centrally-managed of analytics.

. the Big data trend is gererating an enormous amou of information that requires advanced analytics and new market players to take advantage of it.

Criteria For Big Data projects.
1- speed of decision making
2. Through put
3. Analysis Flexibility.

DATE: _____ SUBJECT: lec 2

Three Keys roles of the new Data Ecosystem.

- Deep analytical Talent (Data scientist) — people with advanced training in quantitave disciplines such as mathematics, statistics, machine learning.

- Data sawy professionals (Analysts, Data sawy managers) — people with abasic Knowledge of statistics and/or machine learning who can define key questions that can be answered using advanced analytics.

- Technology & Data Enablers — people providing technical expertise to support analytics projects skill sets including Computer programing and DB administrator.

Data scientist Key Activities

1- reFrame business challenges as analytics challenges
2- Design implement and deploy statistical models and data mining Techniques on big data.
3. Create insights that lead to actionable recommendations.

DATE: _____ SUBJECT: _____

«Data Analytics life cycle»

- Value of using the data Analytics lifecycle.

1- Ensure rigidity and completeness.
2- Enable better transition to members of the cross-Functional analytic teams.

Creating and documenting a process will help demonstrate rigor in your findings.

- repeatable.
- Scale to additional analysts.
- Support validity of findings.

Need for a process to Guide Data Science projects.

1- well-defined processes can help guide any analytic project
2- Focus of Data Analytics ~~project~~ lifecycle is on Data Science projects, not business intellegence.
3. Data science projects tend to require a more consultative approach, and differ from BI projects in a Few ~~ye~~ ways.

- less predictable data
- more projects which lack shape or structure.
- more due diligence in discovery phase.

DATE: _____ SUBJECT: _____

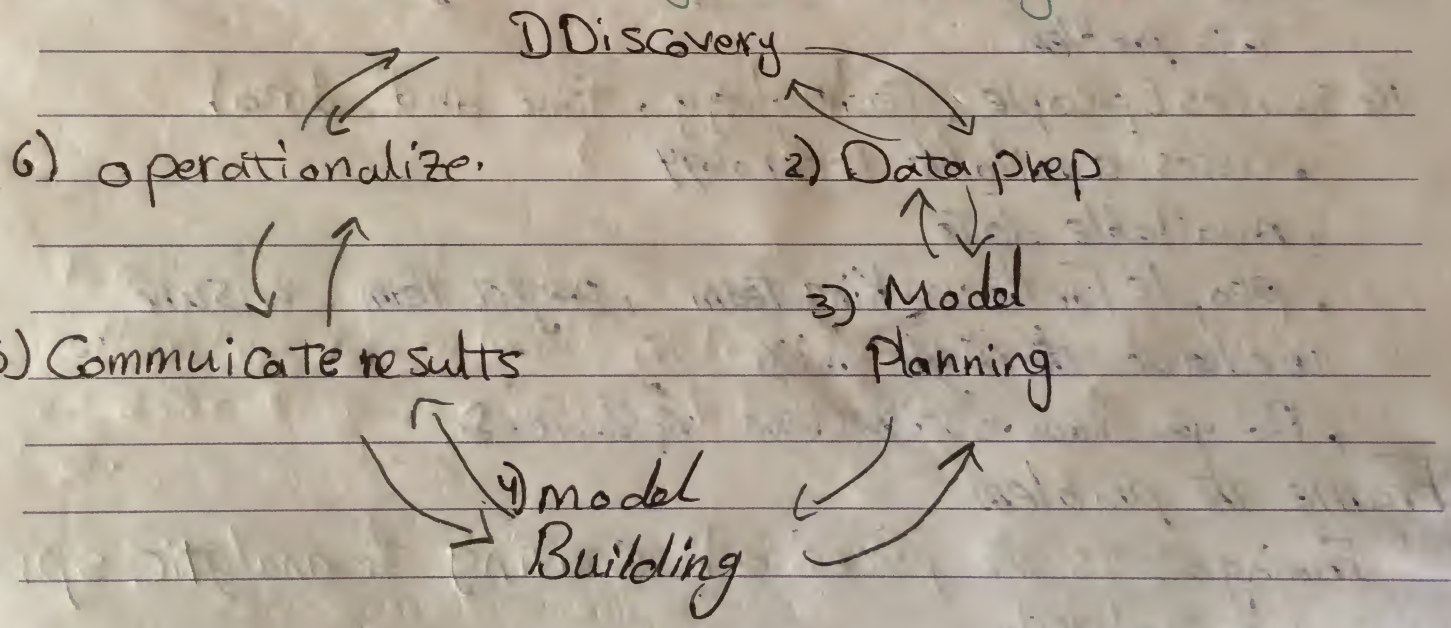
6) Key roles for a Successful Analytic project.

Role	Description
Business user	Someone who benefits From the end results and Can Consult and advise project team on value of end results and how these will be operaTionalized
project sponsor	person responsible For the genesis of the projects providing the motives For the project and Core business problem, generally provide the funding and will measure the degree of value From the Final outputs of the working team.
project manager	Ensure key objectives are met on time and at expected quality.
Business inteligence Analyst	Business domain expertise with deep understanding of the data KPIs, Key metrics and business intelligence From a reporting prespective.
Data Engineer	Deep technical skills to assist with Tuning SQL queries For data management extraction and support data realize to analyti Sand Box

DATE: _____ SUBJECT: _____

Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of working team
Data scientist	provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met

Data Analytics Life Cycle



* Data Analytics Life Cycle :-

We can go back and refine work done in prior phases given new insights and information that you've uncovered

DATE: ____________ SUBJECT: ____________

Phase 1: Discovery.

- learn the Business Domain.
  - determine amount of domain knowledge needed to orient you to data and interpret results down stream.
  - Determine the general analytic problem type "such as clustering, classification"
  - then conduct initial research to learn about the domain area you'll be analyzing.
- learn From the past.
  - learn how previous attempts in the organization to solve this problem.

Resources (people, technology, time and data)
- asses available technology
- Available data
- people for the working team, project team ensure we have the right mix.
- Do you have sufficient resources.

Frame the problem

Framing ⟹ is the process of stating the analytic s problem To be solved.

- state the analytics problem, why it is important

Summary of discoveryphase.

1) learn about domain knowledge.
2) detect available resources.
   (.... - technology - data - Tools ١) زي امكانيات الاستخدام

تحديد نوع learning "clustering أو classification"

ALAQSA

DATE: Lec 4 SUBJECT:

« data preperation phase »

تعتبر أطول وأصعب مرحله.

حيث هنا بتكون هنا Sand BoX وقد تكون في بعض الأحيان أكبر من data ware house.

١- بيأخذ كل data الممكنة عنده.

٢- تكوين Limited data: هنا ينقل SandBox

ويوجد هنا 2 technologes

1) ETL extract Transform Load.

2) ELT extract load Transform.

حيث يتجمع وتنقل الداتا المتاحه وهو يقرر الى هو عايزه «deep analysis»

وهنا يتم التعامل مع DWH , it

٣- check data.

Valid

Concistance

process

missing Fields

1. Check data Type « structured, unstructure, semi structured »,

2. systematic error.

1- prepare Analytic SandBox

2- perform ELT

3- Familiarize yourself with the data thoroughly

4- Data Conditioning

5- Survey & visualize.

6- Determine methods

7- Techniques & workFlow.

DATE: ______ SUBJECT: ______

* phase 3 :- Model planning

1- Data exploration
2- Variable selection
3- Model Selection :

١- تحديد framework بناءً عليها ال Techniq الي هستخدمها في الـ...
٢- workflow, Feature selection الي تعتمد عليه model
٣- تحديد الشكل النهائي لـ model بناءً عليها.

* phase 4 :- model Building
1- develop data sets for testing, Training, and production purposes.
2- get the best environment you can for building models and work flows (R, RapL, ...)

١- نشوف نوع data وهل هتكون مناسبة ولا لأ.
٢- التعديل في ال model لو احتاج الأمر شيء أو بعض الأخطاء.
تحاول important parameters - valid - accuracy, mistakes ما يلكس.

* phase 5 :- Communicate Results

١- أنه نحاول إقناع Sponser بالقرارات التي تم تحديدها سابقاً.
٢- وعرض النتائج عليه.
Did we succeed? Did we fail?

DATE: SUBJECT:

* phase 6 : operationalize

. تشغيل Test لل system عشان check هل مناسب ولا لا

4 Core Deliverables to Meet most stakeholder Needs

1 - presentation for project sponsors

- Big picture takeaways for executive level Stakeholders.
- Determine key messages to aid their decision-making process
- Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.

2 - presentation for Analysis

- Business process changes
- reporting changes
- Follow Data Scientists will want the details and are Comfortable with Technical graphs

3 - Code. For Technical people.

4 - Technical speces of implementing the code.

⇒ Analyst wish list for a successful Analytics project.

* Data & workspaces.

1 - Access all data
2 - up-to-date data dictionary
3 - Ability to move data back between staging
4 - Sand box.

DATE: _____ SUBJECT: _____

Tools:.

- Statistical, mathematical, visual SW.
- Tool or place to log errors with systems.
- Collaboration → online platform for communication with team members.

DATE: Lec. SUBJECT:

estimated	Actual	
+ ve	+ ve	→ Tp
- ve	- ve	→ TN
+ ve	- ve	→ FP
- ve	+ ve	→ FN

$$Accuracy = \frac{TF + TN}{Tp + TN + FN + FP}$$

- which i/p
- Bulding model that solve the problem
- Check accuracy of mode → Compare with myself
  → Compare " others in same problem

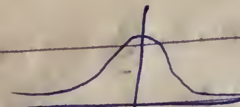
* hypotheis testing وذلك بنفس نتائج دنفس samples يعني زي ما هو
→ Choosing variable → لو حصل ترتيب الأفراد الاغنيا
variance, mean وذلك عدم وجود ← لو منش هيحصل فرق بيتعامل عنه د

الغرض من بناء model هنا ليس تكوين نتائج ولكن من أجل نسبة نتائج موجوده بالفعل.

*median لو فردى يبقى المتصف.
زوجى بنجمع 2 فى النص ونقسم على 2.

وجدوا عند حساب mean إنه هناك احتاليه الحصول على 0
هنا تم تحديدنا انه 2 زي بعض فى حسب قد تكون ± د
زى مختلفة لذلك تم اللجوء الى variance.

Test check difference of mean.



يشترط على normal distribution. 1. t. test الأصلي

2. welchs t. test default of R << more tightly.

3. Rank sun → general << less tightly.

عام استخدامه لأي نوع من الداتا.

H0 ⇒ "hypothes" وتعني مش حدث أي تغير

H1 ⇒ " " " تعني وجود حل جديد

T. test "two way test" الأصلي - inter

1- يشترط ال 2 distribution مش لازم نفس variance.

أما student T. test يشترط ال 2 يكون ليهم نفس variance.

(1)

→ . t = 0 accept null hypotes. مفيش فرق

→ . t = vv area under the curve. في فرق

accept intermediate hypothesis. ← ينتج منه

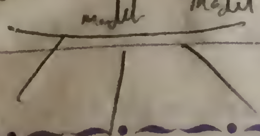
RackSun.

w → old new threshold value.

Samples			A	
X_1	X_2	X_3	X_{11}	X_{21}
			X_{12}	X_{22}

$w = \sum sgn(-old + new)$

model model



X →	0	+	−
sgn(x)	0	1	−1

DATE: ______________ SUBJECT: ______________

power ⟹ Positive الحقيقية أكثر

signif ⟹ F.P error قلة

effect size ⟹ Actual difference between 2 means.

ANOVA

يقدر يشتغل على one way (1 variable) ولا Two way (2 way).

• مش شرط يكونوا عندي 2 model اللي "old" و "new"

• و ممكن يكونوا أكثر من 2 model

S28 | مهم جداً

1. Calculate mean of each population.

①	②	③
1	2	2
2	4	3
5	2	4

$m_1 = 2.67$

$m_2 = 2.67$

$m_3 = 3$

2- grand mean: $m_G = \frac{m_1 + m_2 + m_3}{n} = \frac{2.67 + 2.67 + 3}{3} = 2.78$

2. Sum of Squares (SS)

$$SS_{within} = \sum (X_1 - m_1)^2 + \sum (X_2 - m_2)^2 + \sum (X_3 - m_3)^2$$

(لكل sample)

$(1-2.67)^2 + (2-2.67)^2 + (5-2.67)^2 + (2-2.67)^2 + (4-2.67)$

$$SS_{total} = \sum (X - m_G)^2 = 13.6$$

لكل شيء

SUBJECT:

DATE:

- $SS_{Between} = SS_{total} - SS_{within} = 0.23$

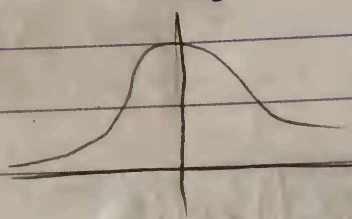
- $S_w^2 = V_w = \frac{SS_w}{N - K} = 13.34/(9-3) = 2.22$

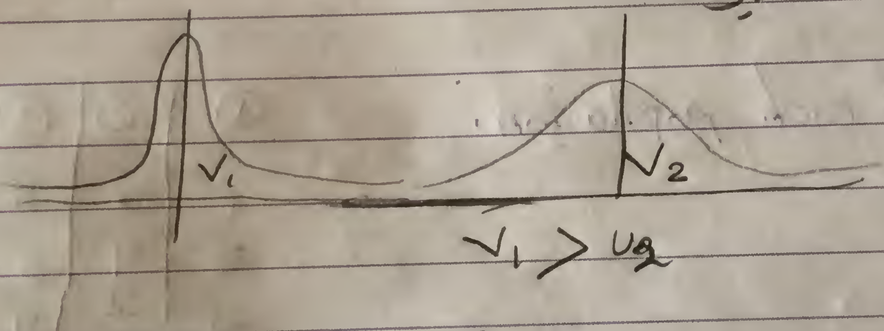
"9" Samples عدد كل ← $N - K$ → "3" models عدد الـ

- $S_B^2 = \frac{SSB}{K-1} = \frac{0.23}{2} = 0.12$

- $F = S_B^2 / S_w^2$ → <1 accept null hypothes.
  → >

S_w → كل عاليكين ← Variance بيكون نطاق أوسع أكبر





$V_1 > V_2$

((ALAQSA))